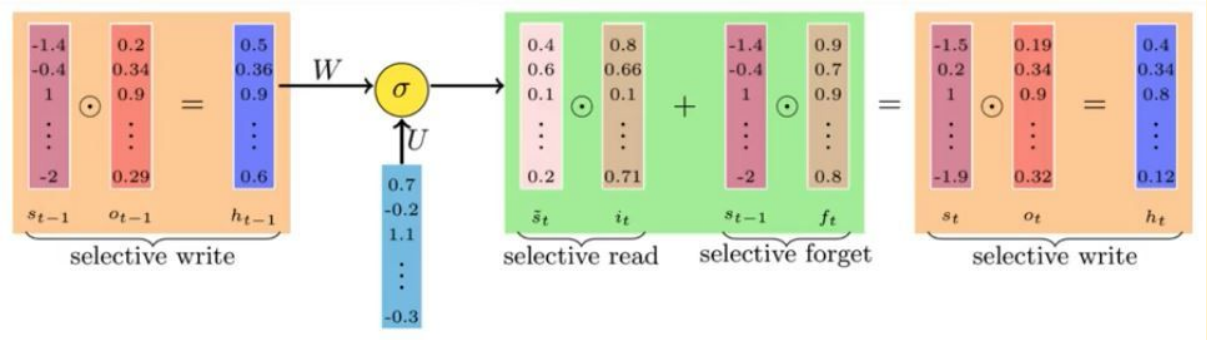
PadhAI: Vanishing and Exploding gradients in LSTM

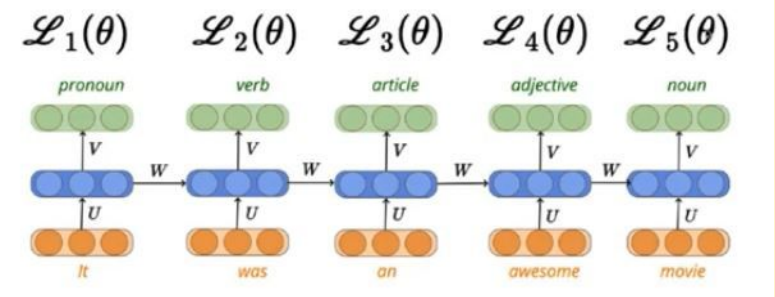
One Fourth Labs

How gates help to solve the problem of vanishing gradients:

- During forward propagation the gates control the flow of information, similarly during back propagation gates control the flow of gradients.
- Gates prevent any irrelevant from being written to the state. It is easy to see that during back propagation gradients will get multiplied by the gate.

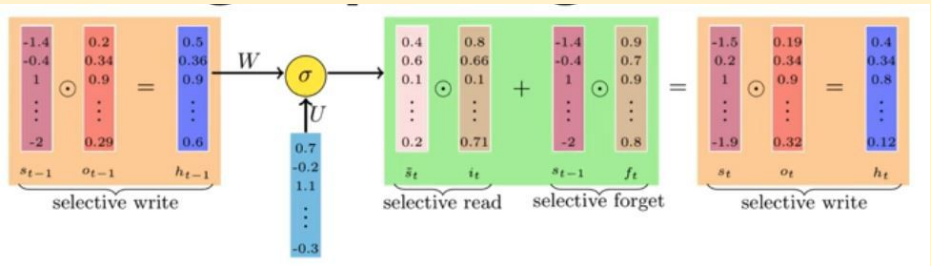


- If the loss at $\mathbf{L_t(\theta)}$ was large because **W** not good enough to complete $\mathbf{s_1}$ correctly then this information will not be propagated back to **W** as the gradient $\frac{\partial Lt(\theta)}{\partial W}$ will vanish.



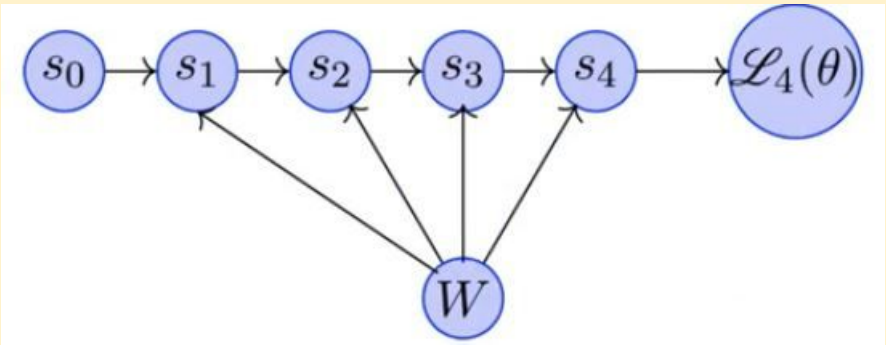
$$\frac{\partial \mathscr{L}_t(\theta)}{\partial W} = \frac{\partial \mathscr{L}_t(\theta)}{\partial s_t} \sum_{k=1}^{t} \prod_{j=k}^{t-1} \frac{\partial s_{j+1}}{\partial s_j} \frac{\partial^+ s_k}{\partial W}$$

- If the state at time **t-1** did not contribute much to the state at time **t** (i.e., if $\mathbf{||f_t||}$ —> **0** and $\mathbf{||o_{t-1}||}$ —> **0**) then during back propagation the gradients flowing into $\mathbf{s_t}$**-1** will vanish.
- The key difference from vanilla RNNs is that the flow of information and gradients is controlled by the gates which ensure that the gradients should not vanish when they should (i.e., when $\mathbf{s_{t-1}}$ didn't contribute much to $\mathbf{s_t}$)



Revisiting vanishing gradients in RNNs:

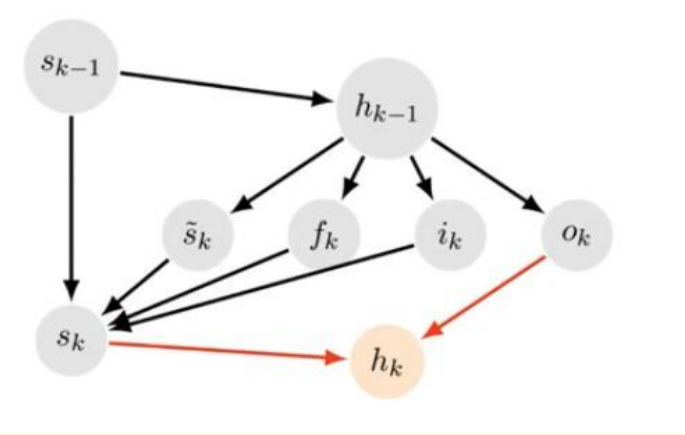
- In general the gradient of $L_t(\theta)$ w.r.t. θ_i vanishes when the gradients flowing through each and every path from $L_t(\theta)$ to θ_i vanish.



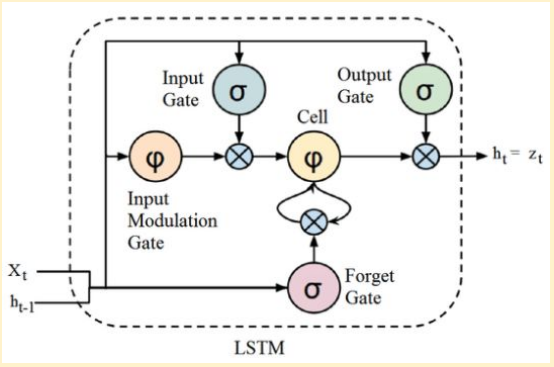
- On the other hand, the gradient of $L_t(\theta)$ w,r,t. θ_i explodes when the gradient flowing through at least one path explodes.
- Here's the nice visualisation for LSTMs for better understanding of concepts.

Visualizing the LSTMs

Dependency diagram for LSTMs:



$$
\begin{aligned}
o_k &= \sigma(W_o h_{k-1} + U_o x_k + b_o) \\
i_k &= \sigma(W_i h_{k-1} + U_i x_k + b_i) \\
f_k &= \sigma(W_f h_{k-1} + U_f x_k + b_f) \\
\tilde{s_k} &= \sigma(W h_{k-1} + U x_k + b) \\
s_k &= f_k \odot s_{k-1} + i_k \odot \tilde{s_k} \\
h_k &= o_k \odot \sigma(s_k)
\end{aligned}
$$



- The **long-term memory** is usually called the **cell state**. The looping arrows indicate recursive nature of the cell. This allows information from previous intervals to be stored with in the LSTM cell. Cell state is modified by the forget gate placed below the cell state and also adjust by the input modulation gate. From equation, the previous cell state forgets by multiply with the forget gate and adds new information through the output of the input gates.

- The **remember vector** is usually called the **forget gate**. The output of the forget gate tells the cell state which information to forget by multiplying 0 to a position in the matrix. If the output of the forget gate is 1, the information is kept in the cell state. From equation, sigmoid function is applied to the weighted input/observation and previous hidden state.

- The **save vector** is usually called the **input gate**. These gates determine which information should enter the cell state / long-term memory. The important parts are the activation functions for each gates. The input gate is a **sigmoid** function and have a range of [0,1]. Because the equation of the cell state is a summation between the previous cell states, sigmoid function alone will only add memory and not be able to remove/forget memory. If you can only add a float number between [0,1], that number will never be zero / turned-off / forget. This is why the input modulation gate has a **tanh** activation function. Tanh has a range of [-1, 1] and allows the cell state to forget memory.

For more detailed explanation, here's a great post on medium by Eugine Kang: LSTM Concepts .

Computing the gradients:

- We are interested in knowing if the gradient flows to $\mathbf{W_f}$ through $\mathbf{s_k}$.
- It is sufficient to show that $\frac{\partial Lt(\theta)}{\partial W}$ does not vanish (because if this does not vanish we can reach $\mathbf{W_f}$ through $\mathbf{s_k}$).

$$t_0 = \frac{\partial \mathcal{L}_t(\theta)}{\partial h_t} \frac{\partial h_t}{\partial s_t} \frac{\partial s_t}{\partial s_{t-1}} \cdots \frac{\partial s_{k+1}}{\partial s_k}$$

$$h_t = o_t \odot \sigma(s_t)$$

Here $\mathbf{t_0}$ is the gradient of $Lt(\theta)$ at time step **0** w.r.t. $\mathbf{h_t}$.

$$\frac{\partial h_t}{\partial s_t} = \mathcal{D}(o_t \odot \sigma'(s_t))$$

Gradient of $\mathbf{h_t}$ (hiddent state at time step **t**) w.r.t. $\mathbf{s_t}$ (cell state at time step t)

For better grip of Gradient of a function: Gradients and Partial Derivatives.

When do the gradient vanishes:

At time step **t = 0**

$$t_0 = \frac{\partial \mathcal{L}_t(\theta)}{\partial h_t} \frac{\partial h_t}{\partial s_t} \frac{\partial s_t}{\partial s_{t-1}} \cdots \frac{\partial s_{k+1}}{\partial s_k}$$

The gradient of loss function w.r.t. hidden state.

as

$$\frac{\partial \mathcal{L}_t(\theta)}{\partial h_t} = \mathcal{L}'_t(h_t),$$

$$\frac{\partial h_t}{\partial s_t} = \mathcal{D}(o_t \odot \sigma'(s_t)),$$

$$\frac{\partial s_t}{\partial s_{t-1}} \cdots \frac{\partial s_{k+1}}{\partial s_k} = \frac{\partial s_t}{\partial s_{t-1}} \cdots \frac{\partial s_{k+1}}{\partial s_k}$$

Then the above equation of $\mathbf{t_0}$ becomes

$$t_0 = \mathcal{L}'_t(h_t).\mathcal{D}(o_t \odot \sigma'(s_t))\mathcal{D}(f_t) \ldots \mathcal{D}(f_{k+1})$$

$$= \mathcal{L}'_t(h_t).\mathcal{D}(o_t \odot \sigma'(s_t))\mathcal{D}(f_t \odot \ldots \odot f_{k+1})$$
$$= \mathcal{L}'_t(h_t).\mathcal{D}(o_t \odot \sigma'(s_t))\mathcal{D}(\odot_{i=k+1}^{t} f_i)$$

Now, if the second term which is the derivative of the hidden state at time step **t**, is very small then other two terms which when it will get multiplied by yields even smaller value for gradients which is known as vanishing gradients. And if the second term is very large the product will also be very large which is known as exploding gradients.

Dealing with exploding gradients:

At time step **t = 1**

$$t_1 = \frac{\partial \mathcal{L}_t(\theta)}{\partial h_t}\left(\frac{\partial h_t}{\partial o_t}\frac{\partial o_t}{\partial h_{t-1}}\right)\ldots\left(\frac{\partial h_k}{\partial o_k}\frac{\partial o_k}{\partial h_{k-1}}\right)$$

The gradient of loss function w.r.t. hidden state.

as

$$\frac{\partial \mathcal{L}_t(\theta)}{\partial h_t} = \mathcal{L}'_t(h_t),$$

$$\left(\frac{\partial h_t}{\partial o_t}\frac{\partial o_t}{\partial h_{t-1}}\right) = (\mathcal{D}(\sigma(s_t) \odot o'_t).W_o)$$

$$\left(\frac{\partial h_k}{\partial o_k}\frac{\partial o_k}{\partial h_{k-1}}\right) = W_o$$

Then the above equation of $\mathbf{t_1}$ becomes

$$t_1 = \mathcal{L}'_t(h_t)\,(\mathcal{D}(\sigma(s_t) \odot o'_t).W_o)\ldots(\mathcal{D}(\sigma(s_k) \odot o'_k).W_o)$$

$$\|t_1\| \leq \|\mathcal{L}'_t(h_t)\|\,(\|K\|\|W_o\|)^{t-k+1}$$

Now, if we see the exponent on final term which is **t-k+1,** which denotes that is $\mathbf{||W_0||}$ is large then the whole term will increase exponentially which yields very large gradients or gradients will explode.